Specific Aims

Many lung diseases are influenced by a complex set of genetic and genomic factors acting in concert to effect biological changes and disease progression. Simple association methods are widely used to analyze the relationship between genetic or genomic factors and phenotypic outcomes. However, these established approaches fail for multiple reasons, including that they do not elucidate the complex relationship between genetics and diseases and exclude potentially valuable information from other genomic data. Methods that allow for both the universal use of integrative approaches and the quantification of their error are essential to gaining complete, reproducible insight into the complex genomic basis of disease.

To adequately represent the complex relationships among genomic data types, mediation and network analysis have been introduced as as methods that can integrate such data. Mediation analyses characterize the structural relationship between a feature and an outcome, where the feature may act directly, through a mediator or through other mechanisms. Networks provide a graphical representation of links between distinct data types. For instance, they can represent the associations between transcription factors and genes in an inferred regulatory network. Each of these approaches have had success in identifying biological features leading to disease onset and progression, but are limited with respect to appropriate data settings and quantification of error.

In this proposal, I develop statistical methods with computational implementation to extend mediation and network approaches for modeling complex relationships that act together to produce phenotypic changes. They will each be applied to lung diseases to provide more insight to the biological pathways leading to disease. The plans to achieve this are summarized in the following specific aims.

AIM 1: To extend mediation methodology within the standard regression framework for the setting of common binary outcomes for integrative genomic analysis.

The objective of this aim is to develop methodology to characterize a mediation relationship with a common binary outcome without imposing additional assumptions on the data. This is oftentimes the case when we have genetic variants as the exposure, which can act either directly or through a mediator such as gene expression or DNA methylation and a common phenotypic or disease outcome. I will obtain data from the Normative Aging Study and assess whether the relationship of smoking and asthma in the elderly is mediated through DNA methylation.

AIM 2: To develop uncertainty measures for network metrics to account for error propagated in intermediate analyses in networks of genomic data.

The objective of this aim is to construct measures of error corresponding to the metrics of scientific interest in network models. I will apply this approach to an eQTL network derived from the COPDGene study, which has collected various genomic data on chronic obstructive pulmonary disease (COPD). I will then construct statistical tests to compare the structural properties of the network from COPDGene to other COPD eQTL networks.

This research will address gaps in the mediation and network literature; gaps that are of increasing relevance as studies obtain more complete genomic data on participants. The integrated approach I propose to develop will allow for new insights into the mechanisms that relate genetic variants to downstream disease. Furthermore, these aims will give me experience in working with multiple genomic data types, computing, collaborating with experts, and translating interdisciplinary methods to interpretable statistical results. This will help me develop the skills necessary to transition from my PhD to ultimately becoming an independent academic researcher in biostatistics with applications to genetics and genomics.